Potts model based on a Markov process computation solves the community structure problem more theoretically and effectively

Hui-Jia Li,[1,2] Yong Wang,[2,3] Ling-Yun Wu,[2,3] Junhua Zhang,[2,3,4] and Xiang-Sun Zhang[2,3,*]

[1]*School of Management Science and Engineering,*
Central University of Finance and Economics, Beijing, 100081, PR China.

[2]*Academy of Mathematics and Systems Science,*
Chinese academy of Science, Beijing, 100190, PR China.

[3]*National Center for Mathematics and Interdisciplinary Sciences,*
Chinese Academy of Sciences, Beijing 100190, China.

[4]*Key Laboratory of Random Complex Structures and Data Science,*
Chinese Academy of Sciences, Beijing 100190, China.

(Dated: March 30, 2015)

Abstract

Potts model is a powerful tool to uncover community structure in complex networks. Here, we propose a new framework to reveal the optimal number of communities and stability of network structure by quantitatively analyzing the dynamics of Potts model. Specifically we model the community structure detection Potts procedure by a Markov process, which has a clear mathematical explanation. Then we show that the local uniform behavior of spin values across multiple timescales in the representation of the Markov variables could naturally reveal the network's hierarchical community structure. In addition, critical topological information regarding to multivariate spin configuration could also be inferred from the spectral signatures of the Markov process. Finally an algorithm is developed to determine fuzzy communities based on the optimal number of communities and the stability across multiple timescales. The effectiveness and efficiency of our algorithm are theoretically analyzed as well as experimentally validated.

PACS numbers:

*Electronic address: zxs@amt.ac.cn

I. INTRODUCTION

Community structure detection [1–3] is a main focus of complex network studies. It has attracted a great deal of attention from various scientific fields. Intuitively, community refers to a group of nodes in the network that are more densely connected internally than with the rest of the network. In the early stage, these studies were restricted to the regular networks. Recently, inspired by several common characteristics of real networks[4], for example the scale-free property, the majority of the studies focus on networks with practical applications. In this meaning, community structure may provide insight into the relation between the topology and the function of real networks and can be of considerable use in many fields.

A well known exploration for this problem is via the modularity concept, which is proposed by Newman et al. [1–3] to quantify a network's partition. Optimizing modularity is effective for community structure detection and has been widely used in many real networks. However, as pointed out by Fortunato et al[5], modularity suffers from the resolution limit problem which concerns about the reliability of the communities detected through the optimization of modularity. In [6], the authors claimed that the resolution limit problem is attributable to the coexistence of multiple scale descriptions of the network's topological structure, while only one scale is obtained through directly optimizing the modularity. In addition, the definition of modularity only considers the significance of the link density from the static topological structure, and it is unclear how the modularity concept based community structure is correlated with the dynamics behavior in the network.

Complementary to modularity concept, many efforts are devoted to understanding the properties of the dynamical processes taken place in the underlying networks. Specifically, researchers have begun to investigate the correlation between the community structure and the dynamics in networks. For example, Arenas et al. pointed out that the synchronization reveals the topological scale in complex networks[7]. In addition, the Markov process on a network was also extensively studied and used to uncover community structure of the network[8]–[11]. In [9][10], the Markov process on a network is introduced to define the distances among network nodes, and an algorithm is then proposed to partition the network into communities based on these distances. In [8], the authors proposed to quantify and rank the network partitions in terms of their stability, defined as the clustered autocovariance in the Markov process taken place on the network.

Potts dynamical model has also been applied to uncover community structure in networks. Detecting community by using Potts model[12], also known as the superparamagnetic clustering method, has been intensively studied since its introduction by Blatt et al[13]. In the model, the Potts spin variables are assigned to nodes of a network with community structure, and the interactions exist between neighboring spins. Then the structural clusters could be recovered by clustering similar spins in the system, which have more interactions inside communities than outside. The physical aspects of the method, such as its dependence on the definition of the neighbors, type of interactions, number of possible states, and size of the dataset, have been well studied[14][15][19]. Reichardt and Bornholdt[16] introduced a new spin glass Hamiltonian with a global diversity constraint to identify proper community structures in complex networks. The method allows one to identify communities by mapping the graph onto a zero-temperature q-state Potts model with nearest-neighbor interactions. Recently, Li et al[17] noticed that a lot of useful information related to community structure can be revealed by Potts model and the spectral characterization. Despite those excellent works, uncovering the dynamic of spin configure across multiple timescales is still a tough task and not yet been clearly answered. In essence, one can consider the time scale as an intrinsic resolution parameter for the partition: over short time scales from the beginning, many small clusters should be coherent; on the other hand as time evolves, there will be fewer and larger communities that are persistent under the dynamic of Potts model. We need to measure the change of the stability or robustness[8] of spin configure as time evolves and furthermore find some reasonable partitions at intermediate timescales. However, using Potts model alone is difficult to solve this problem.

We notice that the dynamics of Markov process can naturally reflect the intrinsic properties of spin dynamics with community structures and exhibit local uniform behaviors. However, the relationship between dynamics of Potts model and the Markov process, has not been well studied. In this work, using the Potts model and spin-spin correlation, we first investigate this phenomenon, and then uncover the relation between community structure of a network and its meta-stability of spin dynamics, and further propose the signature of communities to characterize and analyze the underlying spin configuration. For any given network, one can straightforwardly derive critical information related to its community structure, such as the stability of its community structures and the optimal number of communities across multiple timescales without using particular algorithms. It overcomes

the inefficiency of the classic methods, such as the resolution limitation of Modularity Q [5][18]. Based on the theoretical analysis, we then develop a parameter free algorithm to numerically detect community structure, which is able to identify fuzzy communities with overlapping nodes by associating each node with a participation index that describes node's involvement in each community. We also demonstrate that the algorithm is scalable and effective for real large scale networks.

The outline of the paper is as follows. Section II introduces the Potts model and the motivation of this work. In Section III, we present a Markov stochastic model, which explains the relationship between spectral signatures and community structure. Section IV describes the critical information derived from the model, such as stability across multiple timescales and the optimal number of communities. Our algorithm is formally described in Section V. Then we give some numerical computations for some representative networks to validate the effectiveness and efficiency of the algorithm in Section VI . Finally, Section VII concludes this paper.

II. POTTS MODEL AND SPIN-SPIN CORRELATION

The Potts model is one of the most popular models in statistical mechanics[12]. It models an inhomogeneous ferromagnetic system where each data point is viewed as a marked node in the network. Here the mark is a cluster label, or spin value, associated with the node. The configuration of the system is defined by the interactions between the nodes and controlled by the temperature. At low temperatures, all labels are identical (spins are aligned), which is equivalent to the presence of a single cluster. As temperature rises, the single cluster starts to split and the interactions between weakly coupled nodes gradually vanished.

Consider an unweighted network with N nodes without self-loops, a spin configuration $\{S\}$ is defined by assigning each node i a spin label s_i which may take integer values $s_i = 1, \cdots, q$. Suppose a system of spins can be in q different states. The Hamiltonian $H(S)$ of a Potts model with this spin configuration S is given by:

$$H(S) = \sum_{\langle ij \rangle} J_{ij}(1 - \delta_{s_i s_j}), (i, j = 1, ..., N) \tag{1}$$

where the sum is running over all neighboring nodes denoted as $\langle ij \rangle$, J_{ij} is the interaction

strength between spin i and spin j, and $\delta_{s_i s_j}$ is 1 if $s_i = s_j$, otherwise 0. J_{ij} is set as

$$J_{ij} = J_{ji} = \frac{1}{\langle k \rangle} \exp[-\frac{(d_{ij})^2}{2}], (i, j = 1, ..., N) \tag{2}$$

where $\langle k \rangle$ is the average number of neighbors per node and d_{ij} is the Euclidean distance between nodes i and j. The interaction J_{ij} is a monotonous decreasing function of d_{ij} and the spins s_i and s_j tend to have the same value as d_{ij} becomes smaller if we minimize the $H(S)$.

To characterize the coherence and correlation between two spins, spin-spin correlation function C_{ij} is defined as the thermal average of $\delta_{s_i s_j}$[13–15]:

$$C_{ij} = \langle \delta_{s_i s_j} \rangle \tag{3}$$

It represents the probability that spin variables s_i and s_j have the same value. C_{ij} takes values from the interval [0, 1], representing the continuum from no coupling to perfect accordance of spins i and j. There are two phases in a homogeneous system where J_{ij} is determined. At high temperatures, the system is in the paramagnetic phase and the spins are in disorder. $C_{ij} \approx \frac{1}{q}$ for all nodes i and j, and q is the number of possible spin values. At low temperatures, the system turns into the ferromagnetic phase and all the spins are aligned to the same direction. $C_{ij} \approx 1$ holds for nodes pair i and j.

If the system is not homogeneous but has a community structure, the states are not just ferromagnetic or paramagnetic. We assume that the spins will go through a hierarchy of local uniform states (meta-stable states), as shown in Fig.1, before they reach a globally stable state with all the same value as temperature decreases. In each local uniform state, spin values of nodes within the same communities are identical and the whole system is divided into several different local regions (communities) due to the dense connections. Correspondingly, we can calculate the hitting and exiting time of each local uniform state to analyze its stability. The hitting or exiting time is the timescale that the system just enters or leaves this local uniform state, during which the nodes' spin values will stably stay on this state. In this way we can associate the community structure with a local uniform state. For a well-formed community structure, each community should be cohesive, which means that it is easy for the nodes to hit the local uniform state. Thus, the hitting time should be early. At the same time, communities should stand clear from each other, which means it is hard for nodes to exit the local uniform state, therefore the exiting time should be late.

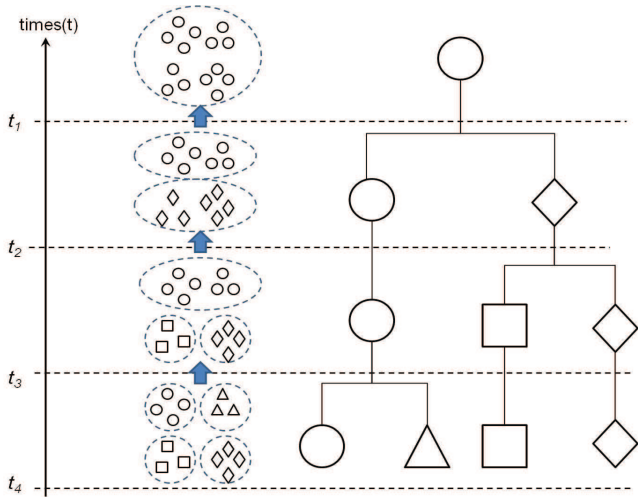


FIG. 1: Dynamics of spin configuration of four communities (A, B, C, D) when they go through several local uniform states to the global stable state with temperature decreasing. Different spin values are described by different shapes. At temperature t_4 ($t_4 > t_3 > t_2 > t_1$, t_i denotes the temperature that i different spin states in the system), we observe four local uniform spin state distributions corresponding to four communities. At temperature t_3, the circle and triangle mix together. At t_2, square with diamond mix together in terms of their hierarchical structure. Finally, at t_1, only one spin state is left, in which all nodes have an identical spin distribution.

Hence, there should be a big gap between the hitting and exiting times when a well-formed community structure exists.

Once J_{ij} has been determined, C_{ij} can be obtained by a Monte Carlo procedure. We used the Swendsen-Wang (SW) algorithm[20] because it exhibits much smaller autocorrelation time[20] than standard methods. For a network with N nodes, the SW algorithm can be briefly described as follows: 1. Generate initial configuration of system $S_1 = (s_1, s_2, .., s_N)$ randomly, where s_i is the spin value of node i randomly chosen from 1 to q, $q = N/2$ is the initial number of spin values. 2. Generate the configuration of system S_2 based on S_1: (a) Visit all pair of nodes $< i, j >$ which have interaction $J_{ij} > 0$, where J_{ij} is the spin interaction computed only based on the adjacent network. Node i and node j are frozen together with probability:

$$p_{ij}^f = 1 - exp(-\frac{J_{ij}}{T}\delta_{s_i,s_j}) \tag{4}$$

where $\delta_{s_i,s_j} = 1$ if $s_i = s_j$ and 0 otherwise. T is the temperature. Calculate all pairs of spins

and put a frozen bond between any frozen pairs. (b) We define SW cluster as the cluster containing all spins that have a path of frozen bonds connecting all of them. Since nodes are frozen only if they have the same spin value, we just need to identify the SW clusters from the same spin values. (c) For each SW cluster, we draw a random number from $1, 2, ..., q$ and assign this number to the values of all nodes of this cluster. After going through all SW clusters, the new configuration S_2 is generated. 3. Iterate Step 2. Then we can calculate the value C_{ij}. We set the initial number of possible spin values $q = N/2$ because if the number of communities is larger than q, some spin states will not be populated. For a specific node, we choose a initial spin value randomly from 1 to q.

III. A STOCHASTIC MODEL

Markov process[26] is a useful tool and has been applied to find communities[8, 9]. In order to establish the connection between the community structure and the local uniform behavior of Potts model, we introduce a Markov stochastic model featured with spectral signatures for the network. Let $A = (V, E)$ denote a network, where V is the set of nodes and E is the set of edges (or links). Consider a Markov random walk process defined on A, in which a random walker freely walks from one node to another along their links. After arriving at one node, the walker will randomly select one of its neighbors and move there. Let $X = X_t, t \geq 0$, denote the walker positions, and $P\{X_t = i, 1 \leq i \leq N\}$ be the probability that the walker hits the node i after exact t steps. For $i_t \in V$, we have $P(X_t = i_t|X_0 = i_0, X_1 = i_1, ..., X_{t-1} = i_{t-1}) = P(X_t = i_t|X_{t-1} = i_{t-1})$. That is, the next state of the walker is determined only by its current state. Hence, this stochastic process is a discrete Markov chain and its state space is V. Furthermore, X_t is homogeneous because of $P(X_t = j|X_{t-1} = i) = p_{ij}$, where p_{ij} is the transition probability from node i to node j.

To relate the Markov process with the patterns of Potts model, p_{ij} is defined as

$$p_{ij} = \frac{C_{ij}}{\sum_{j=1}^{N} C_{ij}} \tag{5}$$

where C_{ij} is the spin-spin correlation function defined in Eq.(4). Via this representation, the tools of stochastic theory and finite-state Markov processes [8][9] can be utilized for the purpose of community structure analysis. Let P be the transition probability matrix, we have:

$$P = D^{-1}C \tag{6}$$

where D is the diagonal degree matrix of C. Let $p_{ij}^{(t)}$ be the probability of hitting node j after t steps starting from node i, we have:

$$p_{ij}^{(t)} = (P^t)_{ij} \tag{7}$$

For this ergodic Markov process, P^t corresponds to the probability of transitions between states over a period of t time steps. To compute the transition matrix P^t, the eigenvalue decomposition of P is used. If λ_k with $k = 1, \cdots, N$ denote the eigenvalues of P, and its right and left eigenvectors u_k and f_k are scaled to satisfy the orthonormality relation[9]:

$$u_k f_l = \delta_{kl}, \tag{8}$$

the spectral representation of P is given by

$$P = \sum_k \lambda_k u_k f_k \tag{9}$$

and consequently

$$P^t = \sum_k \lambda_k^t u_k f_k \tag{10}$$

We assume that eigenvalues of P are sorted such that $\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_N|$. From the theory of spectral clustering[30, 31], P^t can be calculated by a sum of N matrices

$$P^t = \sum_{k=1}^{N} \lambda_k^t \frac{u_n u_n^T D}{u_n^T D u_n} \tag{11}$$

each of which depends only on P's eigensystem. This is accomplished by exploiting the fact that $u_n^T D u_m = I_{nm}$, because P is defined by a normalized symmetric correlation matrix C. Because of the largest eigenvalue $\lambda_1 = 1$, when time $t \to \infty$, $P^{(0)} = P^\infty = \frac{u_1 u_1^T D}{u_1^T D u_1}$. The convergence of every initial distribution to the stationary distribution $P^{(0)}$ corresponds to the fact that the spin of whole system ultimately reaches exactly the same value, as temperature decreases. This perspective belongs to a timescale $t \to \infty$, at which all eigenvalues λ_k^t go to 0 except for the largest one, $\lambda_0^t = 1$. In the other extreme of a timescale $t = 0$, P^t becomes the

stationary distribution matrix. All of its columns are different, and the system disintegrates into as many spin values.

The eigensystem of transition matrix P^t can be naturally correlated with the dynamic process of Potts model. However, it needs preprocessing due to its asymmetrical character. We simply extend P^t to the symmetrical form $G^{(t)} = (P^t + (P^t)^T)/2$ which is defined as the spin correlation matrix at time t. The eigensystem of $G^{(t)}$ have the following correlation corresponding to P^t:

Lemma 1 *The eigenvalues and corresponding eigenvectors of matrix $G^{(t)}$ are exactly same as matrix P^t.*

The proof of lemma 1 is evident. From lemma 1, as $G^{(t)}$ owns the same eigensystem with P^t, it can be used to unfold the dynamic of Potts states. Also, we can use $G^{(t)}$ to find reasonable partitions based on many algorithms, such as the K-means algorithm and GN algorithm[2].

IV. SIGNATURES OF COMMUNITIES IN POTTS MODEL ACROSS MULTIPLE TIMESCALES

In this section, we will uncover the signatures of communities in Potts model across multiple timescales and use this to identify community structure. This scheme benefits from the above analysis, namely the connection between Potts model and Markov process through a stochastic model. A lot of useful information, such as the optimal number of communities, the stability of networks at arbitrary timescale, can be uncovered as follows.

Suppose the partition method divides the network A into K clusters or sets $V_k \subset V, k \in 1, 2, \cdots, K$ which are disjoint and the sets V_1, V_2,..., V_K together form a partition of node set V. The number of nodes in each cluster is denoted by $N_k = |V_k|$. Numerically we will deal with the dynamical process of community structure represented by the spin configuration. We take the time series into consideration. Therefore, we define the the signature of a given community k by the ratio of inner correlations as

$$S_k^{(t)} = \sum_{i,j \in V_k} \frac{[G^{(t)}]_{i,j}}{N_k} \tag{12}$$

$S_k^{(t)}$ can be viewed as a function of timescale t and we can use it to study the trend of community structure as time goes on. Given the number of clusters K, the clusters are found by maximizing the objective function

$$J_K^{(t)} = \sum_{k=1}^{K} \sum_{i,j \in V_K} \frac{[G^{(t)}]_{i,j}}{N_K} \tag{13}$$

over all partitions. The objective can be interpreted as the sum of cluster signature S_k for each cluster V_k. The form of Eq.(13) is related to some famous partition measures, for example, $J_K^{(t)}$ is an extension of the ratio cut criterion defined as the sum of the number of inter-community edges divided by the total number of edges through replacing adjacent matrix A by spin correlation $G^{(t)}$. Furthermore, $J_K^{(t)}$ is also the first part of famous modularity metric Q, which is widely used in the research of community detection.

Further discussion is facilitated by reformulating the average association objective in matrix form. We denote the membership vector of node i by x_i, a probability vector that describes node i's involvement in each community. The element x_i^k means the k-th entry of the membership vector of node i. The hard partition and disjointness of sets V_k requires that the vectors x_i and x_j are orthogonal. The objective $J_K^{(t)}$ can be written in terms of the indicator vectors x_k as

$$J_K^{(t)} = \sum_{k=1}^{K} \frac{x_k^T G^{(t)} x_k}{x_k^T x_k} \tag{14}$$

The objective is to be maximized under the conditions $x_k \in \{0, 1\}$ and $x_i^T x_j = 0$ if $i \neq j$. Eq.(9) can be rewritten as a matrix trace by accumulating the vectors u_k into a matrix $X = (x_1, x_2, ..., x_K)$. We can then write the objective $J_K^{(t)}$ as

$$\begin{aligned} J_K^{(t)} &= tr\{(X^T X)^{-1} X^T G^{(t)} X\} \\ &= tr\{(X^T X)^{-1/2} X^T G^{(t)} X (X^T X)^{-1/2}\} \end{aligned} \tag{15}$$

where matrix $X^T X$ is diagonal. The substitution $Y = X(X^T X)^{-1/2}$ simplifies the optimization problem to $J_K^{(t)} = tr\{Y^T G^{(t)} Y\}$. The condition $Y^T Y = I_K$ is automatically satisfied since

$$Y^T Y = (X^T X)^{-1/2} (X^T X)(X^T X)^{-1/2} = I_K. \tag{16}$$

The vectors y_K thus have unit length and are orthogonal to each other. The optimization problem can be written in terms of the matrix Y as

$$\max_{Y^TY=I} tr\{Y^T G^{(t)} Y\}. \tag{17}$$

Lemma 2 ***(Rayleigh-Ritz theorem)*** *Let L be a symmetric $N \times N$ matrix with eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$ and the corresponding eigenvactors $u_1, ..., u_N$. Then*

$$\max \sum_{k=1}^{K} y_k^T L y_k \quad s.t.\ y_l^T y_k = I \tag{18}$$

equals $\sum_{k=1}^{K} \lambda_k$ and the minimum $y_1, ..., y_K$ lie in the subspace spanned by $u_1, ..., u_K$.

The Rayleigh-Ritz theorem[31] tells us that the maximum for this problem is attained when columns of Y is the eigenvectors corresponding to the K largest eigenvalues of the symmetric correlation matrix $G^{(t)}$. We assume that eigenvalues of P are sorted such that $\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_N|$ and the eigenvector corresponding to λ_k is denoted as u_k. Then the optimal solution of Eq.(18) is the matrix $Y = U = \{u_1, ..., u_K\}$. And the strength of such a cluster is equal to its corresponding t-th power of the eigenvalue

$$S_k^{(t)} = \frac{u_k^T G^{(t)} u_k}{u_k^T u_k} = \lambda_k^t \frac{u_k^T u_k}{u_k^T u_k} = \lambda_k^t \tag{19}$$

For the convergence of the Potts model across multiple timescales, the vanishing of the smaller eigenvalues as the time growing describes the loss of different spin states and the removal of the structural features encoded in the corresponding weaker eigenvectors. For the purpose of community identification, intermediate timescales of local uniform states are of interest. If we want to identify z communities, we expect to find P^t at a given timescale, the eigenvalues λ_k^t may be significantly different from zero only for the range $k = 1, ..., z$. This is achieved by determining t such that $|\lambda_k|^t \approx 0$.

From another perspective, because the eigenvalues are sorted by $\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_N|$, the strength of a community at time t, λ_k^t, can also be viewed as the robustness of k-spin state at time t. At this point, the eigengap $\lambda_k^t - \lambda_{k+1}^t$ can be interpreted as the "difficulty" that the $(k+1)$-spin state transfer to the k-spin state at time t. Given the correlation matrix G, one can measure the most suitable number of possible spins at a specific time t by searching for the value k such that the eigengap $\lambda_k^t - \lambda_{k+1}^t$ is maximized.

The number of communities Λ at time t is then inferred from the location of the maximal eigengap, and this maximal value can be used as a quality measure for the most stable state. The $\Lambda(t)$ is formally defined as

$$\Lambda(t) = arg[max_k(\lambda_k^t - \lambda_{k+1}^t)] \tag{20}$$

From a global perspective if the number of communities Λ is not change for the longest time, we can consider it as the optimal number for this network, represented as Ψ.

The number of communities Λ may keep the same for a long time. However, the variation of spin configuration hidden behind our model is still not clear. To reveal the detail of changes, we need to determine that the timescale of the community structure represented by spin configuration is robust. To a certain extent, the most stable state can represent the spin configuration of the whole network. Thus, we define the stability of community structure at each timescale, $\Theta(t)$, as the stability of the most stable spin state:

$$\Theta(t) = \lambda_{\Lambda(t)}^t - \lambda_{\Lambda(t)+1}^t \tag{21}$$

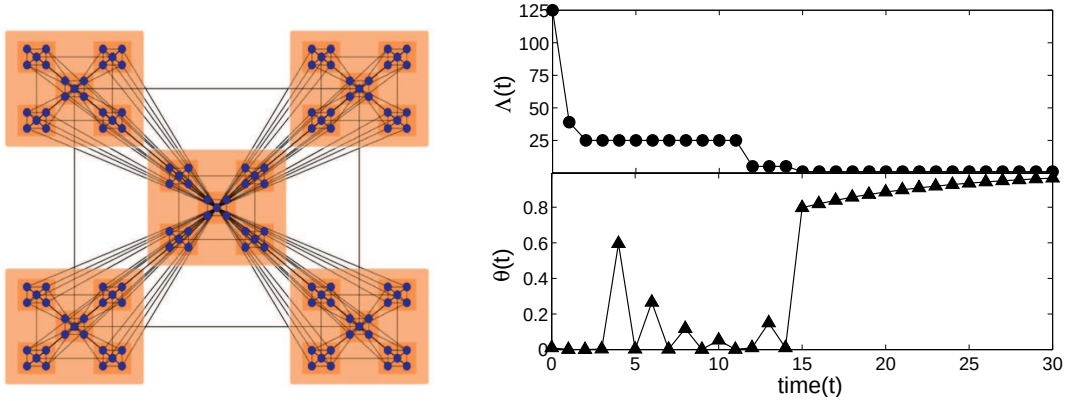
Our expectation is that from the trend of $\Theta(t)$, one can find the most stable timescale for community structure where $\Theta(t)$ reaches the maximal. At this time, the distribution of spin configuration represents the most suitable community structure. Furthermore, from a global perspective, we can use the largest stability corresponding to q communities, $\Gamma(q) = max\{\Theta(t)|\Lambda(t) = q\}$, to indicate the robustness of a network, defined as the stability of the structure with q communities. While $\Gamma(q)$ tries to directly characterize the network structure rather than a specific network partition and thus very convenient to estimate the modularity property of the network.

To show that the model can uncover hierarchical structures in different scales, Fig.2 and Fig.3 give two examples of the multi-level community structures. Fig.2(a) shows the $RB125$ network, which is a hierarchical scale-free network proposed by Ravasz and Barabási in [21]. The regions corresponding to 5 and 25 modules are the most representative in terms of resolution. Next, $H13$-4 proposed by Arenas et al[6] is shown in Fig.3(a), which is a homogeneous degree network with two predefined hierarchical scales. The first hierarchical level consists of 4 modules of 64 nodes and the second level consists of 16 modules of 16 nodes. The partition of both levels are highlighted on the original networks.

In both examples, the most persistent Λ reveals the actual number of hierarchical levels hidden in a network. The signature of such levels can be quantified by their corresponding length of persistent time. The longer the time persists, the more robust the configuration is. From Fig.2(b) and Fig.3(b), we can observe 25 and 16 are the optimal numbers of communities in $RB125$ and $H13$-4 networks owning the longest persistence, respectively. However, 5 modules and 4 modules are also reasonable partitions which show the fuzzy level of the hierarchical networks. These results are in perfect consistence with the generation mechanisms and hierarchical patterns of these two networks.

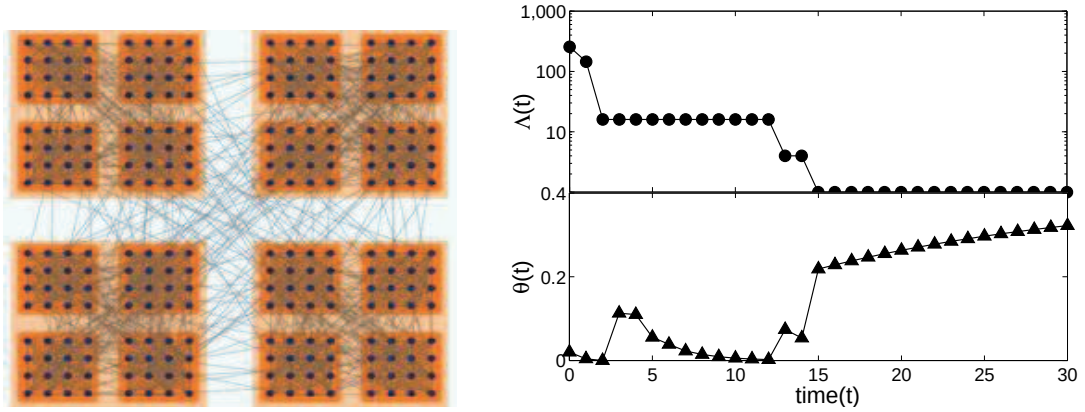
Furthermore, we also show that the variation tendency of stability $\Theta(t)$ in the two cases shed a light on the spin configuration. From Fig.2(b) and Fig.3(b), the corresponding stability $\Theta(t)$ is not a parabolic shape for the timescales of a specific Λ. Thus we cannot easily find the global optimum. However, there are some local maximal values representing better community structure. Thus, we can find these local maximal timescales corresponding to the desirous number of communities and apply G^t to a specific partition method. Furthermore, the stability will reach the lowest value at the end time of all Λ. The stability begins to increase when it transits to new status. One can use $\Theta(t)$ to estimate the modularity property of complex networks, and the larger the Θ the stronger the network community structure. So, one can find the largest corresponding Θ value for a specific number of community Λ and use it to indicate the robustness of modularity structure. For $H13$-4 shown in Fig.3(b), the stability of 16 communities structure, $\Gamma(16) = 0.48$ when $t = 4$, is larger than $\Gamma(4) = 0.31$ when $t = 7$. This indicates that the community structure containing 16 modules is more robust than community structure containing 4 modules. Similarly, for $RB125$ network shown in Fig.2(b), $\Gamma(25) = 0.48$ corresponding to 25 communities structure when $t = 5$ is larger than $\Gamma(5) = 0.31$ when $t = 7$. The robustness of community structure indicated by stability $\Gamma(q)$ favors small but obvious modules. This is the same as [6][7] and is reasonable for many real networks.

Finally, we emphasize the difference between the stability measure proposed in this paper and the modularity Q proposed by Newman[1, 3]. Q is a well-known criterion for evaluating a specific partition scheme of a network. It is defined as "the fraction of edges that fall within communities, minus the expected value of the same quantity if edges fall at random without regard for the community structure" [3, 18, 27, 28]. Different partition schemes will get different Q values for the same network, and larger ones mean better partitions.



(a) (b)

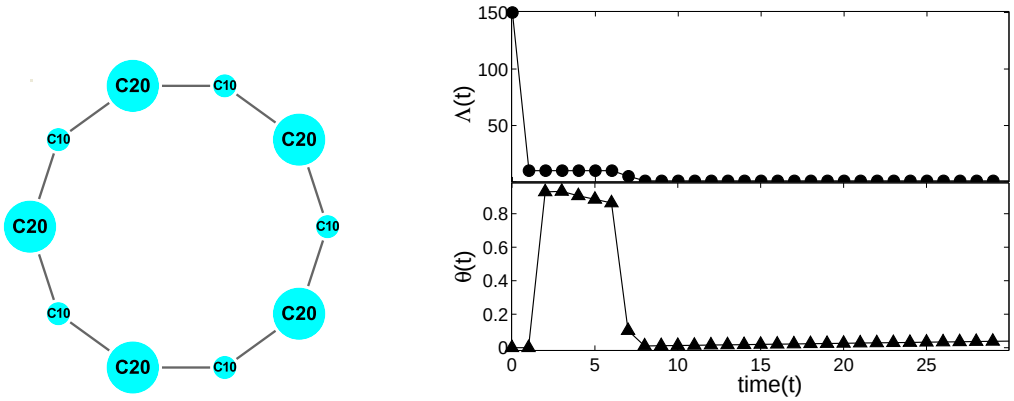
FIG. 2: (a) Structure of $RB125$, with 25 dense communities and 5 sparse communities, are highlighted in the original network. (b) The value of $\Lambda(t)$ and $\Theta(t)$ versus time t.



(a) (b)

FIG. 3: (a) Structure of $H13$-4, with 16 dense communities and 4 sparse communities, are highlighted in the original network. (b) The value of $\Lambda(t)$ and $\Theta(t)$ versus time t.

While our Λ and Γ try to directly characterize and evaluate the structure property based on network's spectra, rather than a specific network partition. Therefore, a network only has exactly self-deterministic Λ and Γ values regardless of how many partition schemes it would have, and the larger the Γ the stronger the network community structure. In addition, Fortunato *et al*[5] pointed out the resolution limit problem of the modularity Q, that is, there exists an intrinsic scale beyond which small qualified communities cannot be detected by maximizing the modularity. However, as shown in Fig.4, when a clique ring contains cliques with different scales (i.e.,the heterogeneous community size), the intrinsic



(a) (b)

FIG. 4: (a) Ring of clique network as a schematic example. Each circle corresponds to a clique, whose size is marked by its label C20 (contains 20 nodes) or C10 (contains 10 nodes). (b) The value of $\Lambda(t)$ and $\Theta(t)$ versus time t.

community structure can be exactly revealed by Λ. With Λ and Γ, we can quantitatively compare the modularity structure of different types of complex networks.

V. A NEW ALGORITHM TO DETECT COMMUNITY

To actually perform the community detection, we propose an approach based on eigenvalue decomposition[29] of correlation matrix $G^{(t)}$. This algorithm allows us to identify multivariate communities across multiple timescales. Based on the above analysis, we correlate the multivariate community structure with the dynamics of the eigenvalues and eigenvectors.

The eigenvalues λ_k and eigenvectors u_k of the symmetric and real-valued matrix $G^{(t)}$ can be obtained by solving the eigenvalue equation

$$G^{(t)} \cdot u_k = \lambda_k^t \cdot u_k, k = 1, ..., N \tag{22}$$

which has N different solutions when time t is small. Assume that the eigenvectors are normalized ($\sum_i u_k(i) = 1$). Each signature $S_k(t) = \lambda_k^t$ is associated with a specific community (the elements in the vector have the same spin value) and quantifies its strength at a given timescale. For each community k, the internal structure is described by the corresponding eigenvector u_k. After normalization ($\sum_i u_k(i) = 1$), its components quantify the relative involvement of each node i to community k by $u_k^2(i)$. Combining the signature

of the community and the index $u_k^2(i)$, the "absolute" involvement of node i in a community k at time t can be described by the following participation index,

$$R_k^{(t)}(i) = \lambda_k^t u_k^2(i) \tag{23}$$

Node i is considered as belonging to community k when the participation index becomes maximal.

From Eq.(23), we observe that participation index evolves as time goes on. When the timescale $t \to \infty$, all eigenvalues λ_k^t approach to 0 except the largest one, $\lambda_0^t = 1$. At this time, all nodes belong to the same community according to the participation index definition. In the other extreme when $t = 0$, the participation matrix R actually becomes the eigenvector matrix U^2. All of its columns are different, and the number of communities is equal to the dimension of the matrix. Here we are interested in the optimal partition at an intermediate timescale with large stability $\Theta(t)$, when the spin configuration represents the most robust community structure. So, we first determine the optimal number of communities by using Λ across long time t. Then, we pick up the timescale t that the stability $\Theta(t)$ is maximal between and $\Lambda(t)$ equals to the optimal number of communities. In many real networks, the formulation of communities is a hard partition and each node belongs to only one cluster after the cluster. This is often too restrictive for the reason that nodes at the boundary among communities share commonalities with more than one community and play a role of transition in many diffusive networks. In our work, the participation index R motivates the extension of the partition to a probabilistic setting. It is extended to the fuzzy partition concept where each node maybe long to different communities with nonzero probabilities at the same time and more reasonable for the real world. Finally, we calculate the participation index at the most stable time t. The framework of the whole process is summarized in Algorithm 1. In the process of the algorithm, calculate the spin-spin correlation matrix C is based on SW algorithm and costs less than $O(N^2)$. It is easy to estimate the computational cost of the algorithm is main on the calculation of eigensystem of G and for sparse graphs, it is about $O(N^2)$. Other steps of the process are some simple matrix computations. So finally, we obtain the cost of Algorithm 1 is $O(N^2)$. Our algorithm is a parameter free method and very easy to implement in real networks.

Algorithm 1 Framework of our new algorithm.

Input:

The adjacent matrix of the network A;

Output:

1: Calculate the spin-spin correlation matrix C.

2: Calculate the Markov transition probability matrix P and G based on C.

3: Calculate the eigenvalues and corresponding eigenvectors of G.

4: Find the optimal number of communities K and corresponding times t with the largest stability.

5: Calculate the participation index R according to Eq.(23).

6: **Return:** Output the participation index R;

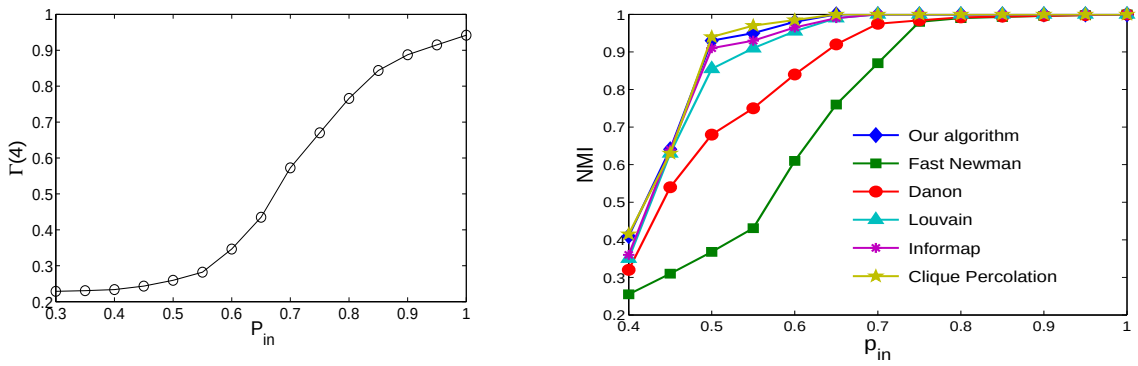
VI. EXPERIMENTS

In this section, we will benchmark the performance of our algorithm. We designed and implemented three experiments for two main purposes: (1) to evaluate the accuracy of the algorithm; (2) to apply it to real large-scale networks.

A. Benchmark network

We empirically demonstrate the effectiveness of our algorithm through comparison with other five well-known algorithms on the artificial benchmark networks. These algorithms include: Newman's fast algorithm[1], Danon et al.'s method[32], the Louvain method[33], Infomap[10], and the clique percolation method[27]. We utilize widely used Ad-Hoc network model, which can produce a randomly synthetic network containing 4 predefined communities and each has 32 nodes. The average degree of nodes is 16, and the ratio of intra-community links is denoted as P_{in}. As P_{in} decreases, the community structures of Ad-Hoc networks become more and more ambiguous, and correspondingly, their $\Gamma(4)$ values climb from 0 to 1, as shown in Fig.5(a).

We use the normalized mutual information (NMI) measure[34] to qualify the partition found by each algorithm. We ask the question whether the intrinsic scale can be correctly uncovered. The experimental results are illustrated in Fig.5(b), where y-axis represents NMI value, and each point in curves is obtained by averaging the values obtained on 50 synthetic networks. As we can see, all algorithms work well when $1 - \mu$ is more than 0.7 with NMI



(a) (b)

FIG. 5: (a) $\Gamma(4)$ values of networks versus different P_{in}. (b) Comparison of accuracy of our algorithm with other five existing algorithms.

larger than 0.85. Compared with other five algorithms, our algorithm performs the best. Its accuracy is only slightly worse than that of the clique percolation when $0.5 \leq 1-\mu \leq 0.65$. However, the complexity of the clique percolation is more than $O(n^3)$ and nearly the same as the time consuming Breadth First Search(BFS). By contrast, the time complexity of our method is very low($O(n^2)$) and can be easily implemented.

B. US Football network

The United States college football team network has been widely used as a benchmark example[1][28] due to its natural community structure. We used the data gathered by Girvan and Newman[1]. It is a representation of the schedule of Division I American Football games in the 2000 season in USA. The nodes in the network represent the 115 teams, while the edges represent 613 games played in the course of the year. The whole network can be naturally divided into 12 distinct groups. As a result, games are generally more frequent between members of the same group than between members of different groups.

First, we calculate Λ and the corresponding stability θ and the results are illustrated in Fig.6. Results show that the optimal number of communities is $\Lambda = 12$, which perfectly agree with the true situation. The stability θ reaches $\Gamma(12) = 0.31$ when $t = 4$. Then we apply our algorithm to the football team network and partitions the network into 12 communities, which is shown in Fig.7. The correct rate of our method is more than 93%, which means that the detected community structure is in a high agreement with the true community

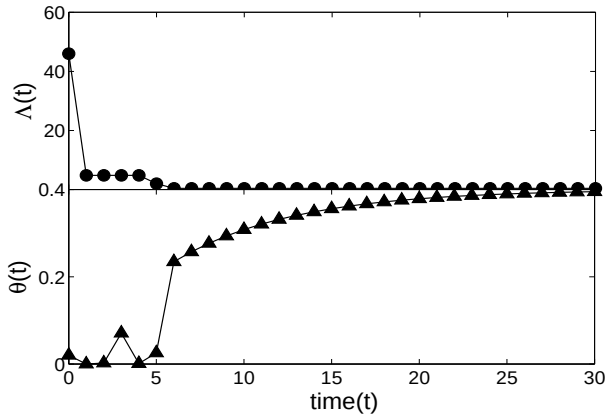


FIG. 6: Computational results of $\Lambda(t)$ and $\Theta(t)$ on US football network.

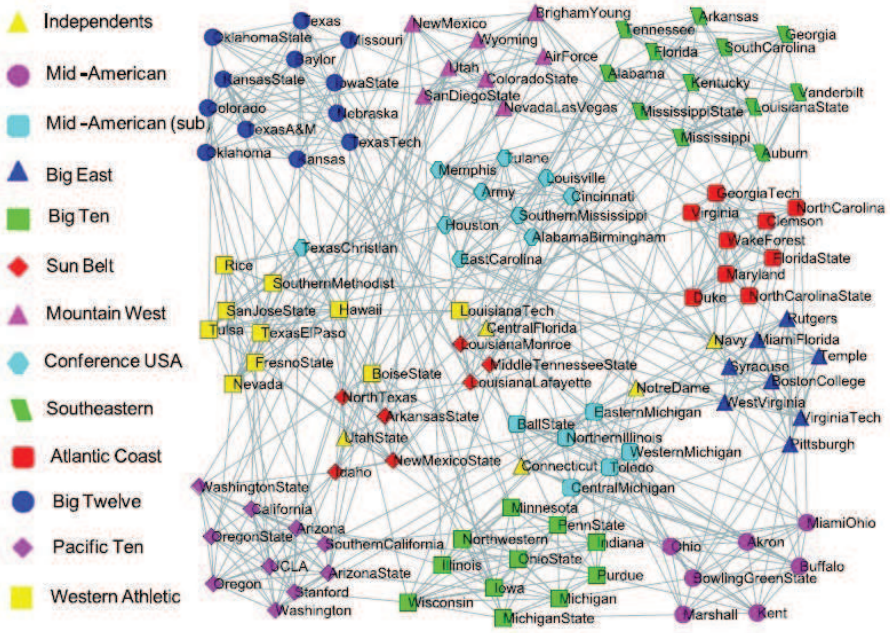


FIG. 7: Computational results of our algorithm on the football team network. The nodes with the same shapes and colors are teams in the same group, and the dense subgraphs in the layout are communities detected by the algorithm. Four fuzzy overlapping nodes are described as independents.

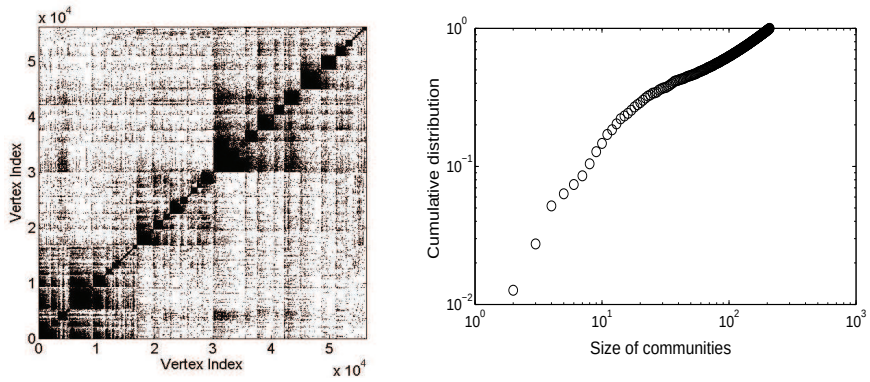
structure. Actually, methods based on optimization of modularity Q usually can just find 11 communities and the correct rate is low due to the fuzziness of the network. We concludes that the ability of our method to reveal a natural characteristic is valuable for many real networks. Furthermore, our algorithm has identified 5 interesting overlapping nodes which described as yellow triangles. The nodes are all fuzzily lie at the boundary communities and

can be viewed as some relative independent clubs which can be interpreted readily by the human eye.

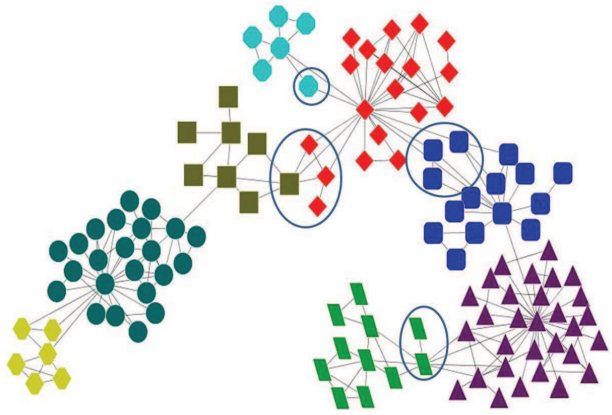
C. Scientific collaboration network

Finally we tested our algorithm on a large-scale network, the scientific collaboration network, collected by Girvan and Newman [22]. The network illustrates the research collaborations among 56,276 physicists in terms of their coauthored papers posted on the Physics E-print Archive at `arxiv.org`. Totally, this network contains 315,810 weighted edges. For visualization purpose, our algorithm outputs a transformed adjacency matrix (in which the nodes within the same community are grouped together) with a hierarchical community structure. From the transformed matrix of Figs.8(a), one can observe a quite strong community structure, or a group-oriented collaboration pattern. Among these physicists, three biggest research communities are self organized regarding to three main research fields: condensed matter, high-energy physics (including theory, phenomenology and nuclear), and astrophysics.

The cumulative distribution of community sizes in power plot is shown in Fig.8(b) and it is a typical scale-free distribution which exists broadly in real world. In total, 737 communities were detected by the optimal community stability, the maximum size of those communities is 195, the minimum size is 2, and the average size is 76. Among these communities, 1,433 of 6,931 pairs of communities have fuzzy participation index with each other. 5% largest communities contain 25.4% of the nodes, while the others are relatively small. The three largest communities correspond closely to research subareas. The largest is solid-state physics, the second largest is super-nuclear physics, and the third is theoretical astrophysics. Furthermore, a subnetwork including eight communities in is shown in Fig.8(c) and four regions including 10 overlapping nodes are highlighted by four circles, which were detected according to the participation index R. The partition result is completely the same as the results in Refs.[22] and [28]. The efficient performance in large real network indicates that our method is useful for further researches in various fields.



(a) (b)



(c)

FIG. 8: (a) Transformed adjacency matrix of the scientific collaboration network. (b) Distribution of community sizes in a linear plot. (c) Subnetwork including eight communities illustrated in different shapes and colors and 10 overlapping nodes enclosed by four circles.

VII. CONCLUSION

In summary, we have presented a more theoretically-based community detection framework which is able to uncover the connection between network's community structures and spectrum properties of Potts model's local uniform state. We demonstrate that important information related to community structures can be mined from a network's spectral signatures through a Markov process computation, such as the stability of modularity structures and the optimal number of communities. Based on theoretical analysis, we further developed an algorithm to detect fuzzy community structure. Its effectiveness and efficiency have been demonstrated and verified through both the simulated networks and the real large-scale networks.

Acknowledgments: We are grateful to the anonymous reviewers for their valuable suggestions which are very helpful for improving the manuscript. The authors are separately supported by NSFC grants 11131009, 60970091, 61171007, 91029301, 61072149, 31100949, 61134013 and grants kjcx-yw-s7 and KSCX2-EW-R-01 from CAS.

---

[1] M.E.J.Newman, Phys. Rev. E. **69**, 066133(2004).

[2] M.E.J.Newman, M.Girvan, Phys. Rev. E. **69**, 026113(2004).

[3] M.E.J.Newman, Proc. Natl. Acad. Sci **103**, 8577-8582(2006).

[4] A.L.Barabási, R.Albert, Science **286**, 509-512(1999).

[5] S.Fortunato, M.Barthelemy, Proc. Natl. Acad. Sci **104**, 36(2007).

[6] A.Arenas, A.Fernandez, S.Gomez, New. J. Phys **10**, 053039(2008).

[7] A.Arenas, A.Diaz-Guilera, C.J.Perez-Vicente, Phys. Rev. Lett **96**, 114102(2006).

[8] J.C.Delvenne, S.N.Yaliraki, M.Barahona, Proc. Natl. Acad. Sci **107(29)**, 12755-12760(2010).

[9] W.N.E, T.Li, E.Vanden-Eijnden, Proc. Natl. Acad. Sci **105**, 7907-7912(2008).

[10] M.Rosvall, C.T.Bergstrom, Proc. Natl. Acad. Sci **105(4)**, 1118-1123(2008).

[11] H.Zhou, Phys. Rev. E **67**, 041908(2003).

[12] F.Y.Wu, Rev. Mod. Phys **54(1)**, 235-268(1982).

[13] M.Blatt, S.Wiseman, E.Domany, Phys. Rev. Lett **76**, 3251-3255(1996).

[14] S.Wiseman, M.Blatt, E.Domany, Phys. Rev. E **57)**, 3767-3783(1998).

[15] H.Agrawal, E.Domany, Phys. Rev. Lett **90**, 158102(2003).

[16] J.Reichardt, S.Bornholdt, Phys. Rev. Lett **93**, 218701(2004).

[17] H.J.Li, Y.Wang, L.Y.Wu, Z.P.Liu, L.Chen, X.S.Zhang, Eur. Phys. Lett **97**, 48005(2012).

[18] X.S.Zhang, R.S.Wang, Y.Wang, J.Wang, Y.Qiu, L.Wang, L.Chen, Eur. Phys. Lett **87**, 38002(2009).

[19] T.Ott, A.Kern, W.Steeb, R.Stoop, J. Stat. Mech **11**, 11014(2005).

[20] S.Wang, R.H.Swendsen, Physica(Amsterdam) **167A**, 565(1990).

[21] E.Ravasz, A.L.Barabási, Phys. Rev. E **67**, 026112(2003).

[22] M.Girvan, M.E.J.Newman, Proc. Natl. Acad. Sci **99**, 7821-7826(2002).

[23] J.Shi, J.Malik, IEEE Tans.On Pattern Analysis and Machine Intelligent **22(8)**, 888-904(2000).

[24] M.Fiedler, Algebraic Connectivity of Graphs. Czechoslovakian Math J **23**, 298-305(1973).

[25] R.Guimera, L.A.N.Amaral, Nature **2**, 895-900(2005).

[26] B.D.Hughes, Random walks and random environments: Random walks, Clarendon Press, Oxford, UK **1**, (1995).

[27] G.Palla, I.Derényi, I.Farkas, T.Vicsek, Nature **435**, 814-818(2005).

[28] Z.P.Li, S.H.Zhang, R.S.Wang, X.S.Zhang, L.Chen, Phys. Rev. E **77**, 036109(2008).

[29] C.Allefeld, M.Muller, J.Kurths, J, Int. J. Bifurcat. Chaos **17**, 3493(2007).

[30] J.Shi, J.Malik, IEEE Trans.Pattern Anal. Mach. Intell **22**, 8888(2000).

[31] A.Azran and Z.Ghaharmani, IEEE Computer Society Conference on Computer Vision and Pattern Recognition **Vol. I**, 190-197(2006).

[32] L.Danon, J,Duch, D.Guilera, A.Arenas, J. Stat. Mech **29**, P09008(2005).

[33] V.D.Blondel, J.L.Guillaume, R.Lambiotte, E.Lefebvre, J. Stat. Mech **10**, P10008(2005).

[34] A.Lancichinetti, S.Fortunato, Phys. Rev. E **80**, 056117(2009).